Remote Sensing of Environment 334 (2026) 115206

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Enhancing cloud detection across multiple satellite sensors using a combined Swin Transformer and UPerNet deep learning model

Shulin Pang [a], Zhanqing Li [b,*], Lin Sun [c], Biao Cao [a], Zhihui Wang [d], Xinyuan Xi [e], Xiaohang Shi [f], Jing Xu [g], Jing Wei [h,*]

[a] Innovation Research Center of Satellite Application, Faculty of Geographical Science, Beijing Normal University, Beijing, China
[b] Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
[c] College of Geomatics, Shandong University of Science and Technology, Qingdao, China
[d] University of Science and Technology of China, Hefei, China
[e] College of Marine Technology, Ocean University of China, Qingdao, China
[f] State Key Laboratory of Climate System Prediction and Risk Management, School of Atmospheric Physics, Nanjing University of Information Science and Technology, Nanjing, China
[g] School of Geography, Nanjing Normal University, Nanjing, China
[h] MEEKL-AERM, College of Environmental Sciences and Engineering, Institute of Tibetan Plateau, and Center for Environment and Health, Peking University, Beijing, China

ARTICLE INFO

Editor : Dr. Menghua Wang

Keywords:
Cloud detection
Cross-sensor
STUPmask
Swin Transformer
UPerNet

ABSTRACT

Cloud detection is crucial in many applications of satellite remote sensing data. Traditional cloud detection methods typically operate at the pixel level, relying on empirically tuned thresholds or, more recently, machine learning classification schemes based on training datasets. Motivated by the success of the Transformer with its self-attention mechanism and convolutional neural networks for enhanced feature extraction, we propose a new encoder-decoder method that captures global and regional contexts with multi-scale features. This new model takes advantage of two advanced deep-learning techniques, the Swin Transformer and UPerNet (named STUPmask), demonstrating improved cloud detection accuracy and strong adaptability to diverse imagery types, spanning spectral bands from visible to thermal infrared and spatial resolutions from meters to kilometers, across a wide range of surface types, including bright scenes such as ice and desert, globally. Training and validation of the STUPmask model are conducted using data obtained from the Landsat 8 and Sentinel-2 Manually Cloud Validation Mask datasets on a global scale. STUPmask accurately estimates cloud amount with a marginal difference against reference masks (0.27 % for Landsat 8 and −0.81 % for Sentinel-2). Additionally, the model captures cloud distribution with a high overall classification accuracy (97.51 % for Landsat 8 and 96.27 % for Sentinel-2). Notably, it excels in detecting broken, thin, and semi-transparent clouds across diverse surfaces, including bright surfaces like urban and barren lands, especially with acceptable accuracy over snow and ice. These encompass the majority of challenging scenes encountered by cloud identification methods. It also adapts to cross-sensor satellite data with varying spatial resolutions (4 m–2 km) from both Low-Earth-Orbit (LEO) and Geostationary-Earth-Orbit (GEO) platforms (including GaoFen-2, MODIS, and Himawari-8), with an overall accuracy of 94.21–97.11 %. The demonstrated successes in the automatic identification of clouds with a variety of satellite imagery of different spectral channels and spatial resolutions render the method versatile for a wide range of remote sensing studies.

1. Introduction

Clouds are ubiquitous, covering approximately 60–70 % of the Earth, particularly over oceans and tropical regions (Asner, 2001; King et al., 2013; Y. C. Zhang et al., 2004). They can obstruct surface observations from space, posing a great challenge in the retrieval of both atmospheric and land surface parameters (B. Li et al., 2021; Schneider et al., 2010; Wang et al., 2023; Wei et al., 2023, 2024; Zhen et al., 2023). However,

* Corresponding authors.
E-mail addresses: zhanqing@umd.edu (Z. Li), jingwei@pku.edu.cn (J. Wei).

https://doi.org/10.1016/j.rse.2025.115206
Received 22 April 2025; Received in revised form 20 November 2025; Accepted 14 December 2025
Available online 9 January 2026

Table 1
The spectral bands of the sensors on Landsat 8 and Sentinel-2 satellites.

Landsat 8			Sentinel-2			Band Type
Band	Wavelength (μm)	Resolution (m)	Band	Wavelength (μm)	Resolution (m)	
1	0.435–0.451	30	1	0.433–0.453	60	Coastal
2	0.452–0.512	30	2	0.458–0.523	10	Blue
3	0.533–0.590	30	3	0.543–0.578	10	Green
4	0.636–0.673	30	4	0.650–0.680	10	Red
–	–	–	5	0.698–0.713	20	Red edge
–	–	–	6	0.733–0.748	20	Red edge
–	–	–	7	0.773–0.793	20	Red edge
5	0.851–0.879	30	8	0.785–0.900	10	NIR
–	–	–	8a	0.854–0.875	20	Red edge
–	–	–	9	0.935–0.955	60	Water vapor
9	1.363–1.384	30	10	1.360–1.390	60	Cirrus
6	1.566–1.651	30	11	1.565–1.655	20	SWIR-1
7	2.107–2.294	30	12	2.100–2.280	20	SWIR-2
8	0.503–0.676	15	–	–	–	Panchromatic
10	10.60–11.19	100	–	–	–	TIRS-1
11	11.50–12.51	100	–	–	–	TIRS-2

the probability of cloud presence depends on satellite overpass time and pixel size (Zhu and Woodcock, 2012, 2014). The complexity of cloud inhomogeneity, morphology, and interactions with underlying surfaces further complicates detection, especially over bright surfaces (Li and Leighton, 1991). Cloud identification has thus been crucial in satellite-based Earth Observation (Arvidson et al., 2001; Irish, 2000).

The exponential growth in the volume of satellite data would require laborious, time-intensive, and costly human resources (Li et al., 2016; Tamiminia et al., 2020). As a result, many methods have been developed and implemented to automatically identify clouds, mostly using empirically tuned thresholds, chiefly due to their simplicity and reasonable accuracy, such as those used for the Advanced Very High Resolution Radiometer (AVHRR) (Kriebel et al., 1989; Saunders and Kriebel, 1988; Stowe et al., 1991) and the Moderate Resolution Imaging Spectroradiometer (MODIS) (Ackerman et al., 1998; Frey et al., 2008). For higher-spatial-resolution sensors, Zhu and Woodcock (2012), for example, introduced an Fmask algorithm, which incorporates many spectral tests to differentiate the spectral characteristics between cloudy and cloud-free scenes for use with Landsat and Sentinel-2 imagery (Qiu et al., 2017, 2019; Zhu et al., 2015). Sun et al. (2016) devised a dynamic threshold algorithm for Landsat 8 imagery based on the mixed pixel decomposition theory and radiative transfer modeling with a priori surface reflectance model database. Frantz et al. (2018) enhanced the accuracy of cloud detection in Sentinel-2 imagery by using the parallax effect to distinguish clouds from bright surfaces in the potential cloud pixels. Such methods are based on spectral differences between cloudy and clear images of the same area over time with different accuracies (Frantz et al., 2015; Gómez-Chova et al., 2017; Hagolle et al., 2010; Jin et al., 2013; Zhu and Woodcock, 2014). Despite their many advantages, the threshold methods suffer from some common limitations, e.g., they often fail over bright areas like barren and snow-covered surfaces due to the low contrast with bright clouds. In addition, multi-temporal approaches require time series images with cloud-free pixels, which can be difficult to obtain in regions frequently obscured by clouds.

In recent years, data-driven artificial intelligence methods have improved cloud detection, leveraging their remarkable data mining capability to extract valuable insights from vast amounts of input features (Pérez-Suay et al., 2018). Particularly, "pixel-level" machine-learning (ML)-based models have advanced substantially from sensors lacking specific spectral channels, e.g., decision trees (Hollstein et al., 2016; Scaramuzza et al., 2012), neural networks (Hughes and Hayes, 2014), Bayesian (Hollstein et al., 2016), support vector machines (Sui et al., 2019), and random forests (Ghasemian and Akhoondzadeh, 2018; Wei et al., 2020). The improved performance of "pixel-level" ML-based algorithms stems from their capacity to iteratively optimize extracted features and identify the most suitable classifier (Jeppesen et al., 2019). Nonetheless, the process of feature selection often depends significantly on manual intervention and operates on a point-wise basis, lacking the ability to incorporate contextual and global information. Deep learning (DL) models, particularly "image feature-based" ML models, such as Convolutional Neural Networks (CNNs), can integrate spectral and spatial information concurrently and have been extensively employed, especially for image classification and object detection (Cheng et al., 2016; Deng et al., 2018). CNN architectures have demonstrated success in cloud detection tasks because they can take advantage of the contrast between the spatial variability of clouds and that of the underlying surface. The Deep Pyramid Network (Ozkan et al., 2018), SegNet (Chai et al., 2019), U-Net (Jeppesen et al., 2019; Wieland et al., 2019; Wright et al., 2024; H. K. Zhang et al., 2024), and Multi-scale Convolutional Feature Fusion (MSCFF) (Z. Li et al., 2019) architectures have demonstrated effectiveness in producing cloud masks that closely resemble manual annotations in cloud detection. The CNN models exhibit robust generalization capabilities and are resistant to overfitting through the combination of regularization techniques (Zheng et al., 2018). However, the CNN network weights are static and lack the ability to adjust dynamically to input variations. Furthermore, current research indicates that CNN models encounter challenges in capturing long-range dependencies and the global context because they are limited by their relatively small receptive fields and have difficulty in integrating distant pixels across the entire image (Luo et al., 2016; Xie et al., 2021).

Transformer, a new generation of powerful DL framework, is getting popular for enhancing the extraction of global image features through its self-attention mechanism (Vaswani et al., 2017). It has been applied to cloud detection in satellite imagery (Singh et al., 2023). Several Transformer-based models have been applied aimed at improving cloud detection performance, such as Vision Transformers (Fan et al., 2024; B. Zhang et al., 2023) and Swin Transformer (Tan et al., 2023), as well as hybrid models combining Transformers with CNNs (Gong et al., 2023; Zhang et al., 2022). However, most previous studies primarily trained models separately for individual sensors, limiting their generalizability. Recently, Wright et al. (2025) developed a deep-learning Omni-CloudMask method for cross-sensor cloud and cloud-shadow detection using dynamic *Z*-score normalization and mixed-resolution training across Landsat-8, Sentinel-2, and PlanetScope. However, it mainly focuses on specific sensor pairs and does not fully address the substantial differences in spatial resolution, spectral characteristics, and orbital configurations across sensors.

To address these issues, our study establishes a comprehensive experimental cloud detection framework that integrates the Swin Transformer (Liu et al., 2021) as the encoder and the Unified Perceptual Parsing Network (UPerNet, Xiao et al., 2018) as the decoder. This framework improves cloud detection performance by leveraging both

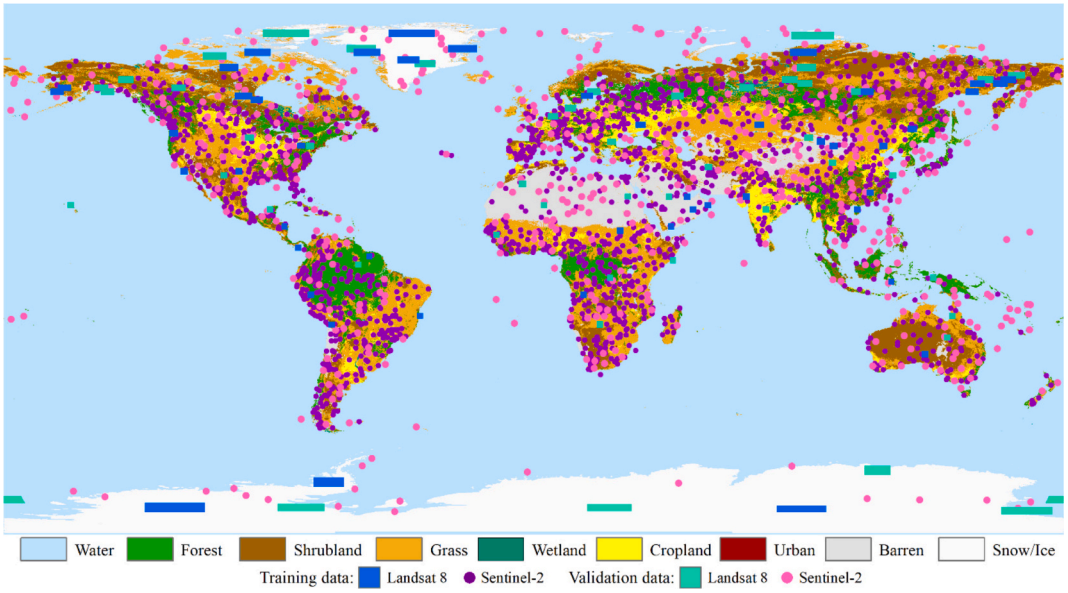

Fig. 1. Geolocation of the global Landsat 8 and Sentinel-2 cloud mask training and validation datasets. The background map displays the MODIS land cover product.

global and regional contexts, as well as multi-scale features for complex scene segmentation, while adapting to variations across different image datasets. More importantly, our STUPmask model is initially pre-trained and applied to two representative high-resolution satellites, Landsat 8 (30 m) and Sentinel-2 (10 m). The model is then evaluated using independent validation and test datasets, which are not used during training, over various underlying surfaces. We further extend the model to accommodate satellites with varying spatial resolutions, from very-high to moderate, across both Low Earth Orbit (LEO) and Geostationary Earth Orbit (GEO) platforms, including GaoFen-2 PMS (4 m), Aqua MODIS (1 km), and Himawari-8 AHI (2 km). Through extensive experiments spanning multiple Earth observation platforms, we provide a practical and reproducible benchmark for cross-sensor generalization. This approach offers valuable insights into how large-scale pretraining can support the development of robust, cross-sensor cloud detection models.

2. Data source

2.1. Landsat and Sentinel imagery

The NASA Landsat series of satellite instruments provide over five decades of high-resolution (~30 m) continuous records of Earth's land surfaces. This dataset is crucial for various applications, including agriculture, forest cover assessment, water resource management, and tracking urban expansion. Currently, Landsat 8 and Landsat 9 are in operation, each carrying two sensors: the Operational Land Imager and the Thermal Infrared Sensor. These sensors provide eleven channels ranging from 0.435 to 12.51 μm, covering visible, near-infrared (NIR), shortwave infrared (SWIR-1 and SWIR-2), and thermal infrared (TIRS) spectra bands, as listed in Table 1 for Landsat 8. This comprehensive coverage allows for ground-based imaging across a wide spectral range. The spatial resolution of the imagery is 30 m, except for the Panchromatic (15 m) and TIRS (100 m) bands (Table 1). Similarly, Sentinel-2 (including 2A and 2B) plays a crucial role in monitoring Earth's land surfaces for various environmental and land management applications. Equipped with MultiSpectral Instrument (MSI) sensors, it has 13 spectral bands spanning from the visible to the SWIR at three high spatial resolutions (10 m, 20 m, and 60 m). Unlike Landsat, Sentinel-2 does not include thermal infrared and panchromatic bands. However, it is one of the few multispectral satellites to include three bands in the red-edge range (Table 1), significantly enhancing its capacity for monitoring vegetation health and related information (Fernández-Manso et al., 2016).

2.2. Training and validation datasets

In this study, the Landsat 8 Biome Cloud Validation Mask (L8 Biome), the Sentinel-2 CloudSEN12, and the Sentinel-2 Cloud Mask Catalogue (S2 CMC) datasets are employed to train the DL model and validate our cloud detection results. The L8 Biome includes 96 globally distributed scenarios (spatial resolution = 30 m) covering a variety of land-use types (Foga et al., 2017), and each Cloud Mask is categorized into three classes based on the percentage of cloudy pixels in the imagery: Clear (less than 35 %), MidClouds (between 35 % and 65 %), and Cloudy (more than 65 %). Additionally, for comparison, we collect Landsat 8 official cloud mask products, which apply spectral reflectance and brightness temperature tests to detect clouds (Foga et al., 2017; Zhu and Woodcock, 2012). It is recorded in the Quality Assessment (QA) band in a 16-bit binary format; specifically, cloud and cirrus information labeled with medium and high confidence, located within the 12th to 15th bits, is used for analysis.

CloudSEN12 is a large dataset designed for cloud semantic understanding, comprising 9880 regions of interest (ROIs) and 49,400 image patches (IPs) distributed across all continents except Antarctica, and provides separate training, validation, and testing datasets (Aybar et al., 2022). Each IP spans 5090 × 5090 m and includes data from Sentinel-2 levels 1C and 2A, along with annotations for thick and thin clouds, cloud shadows, Sentinel-1 SAR, digital elevation models, surface water occurrence, land cover types, and cloud mask results from six advanced cloud detection algorithms. Each ROI contains five 5090 × 5090-m patches captured on different dates, corresponding to various cloud cover categories: clear, low-cloudy, almost clear, mid-cloudy, and cloudy. The S2 CMC dataset comprises 513 sub-scenes, each with an image size of 1022 × 1022 pixels (Francis et al., 2020), evenly distributed across 11 surface types worldwide.

To ensure that the training and validation samples encompass nearly all cloud types as well as diverse land-cover types, satellite scenes over various underlying surfaces are selected. For Landsat 8, a total of 48 scenes from the 96 Biome images are chosen for training using the stratified sampling method, which ensures a uniform selection of images and representation of cloud cover across various land-cover types, providing a more balanced and representative dataset for model training

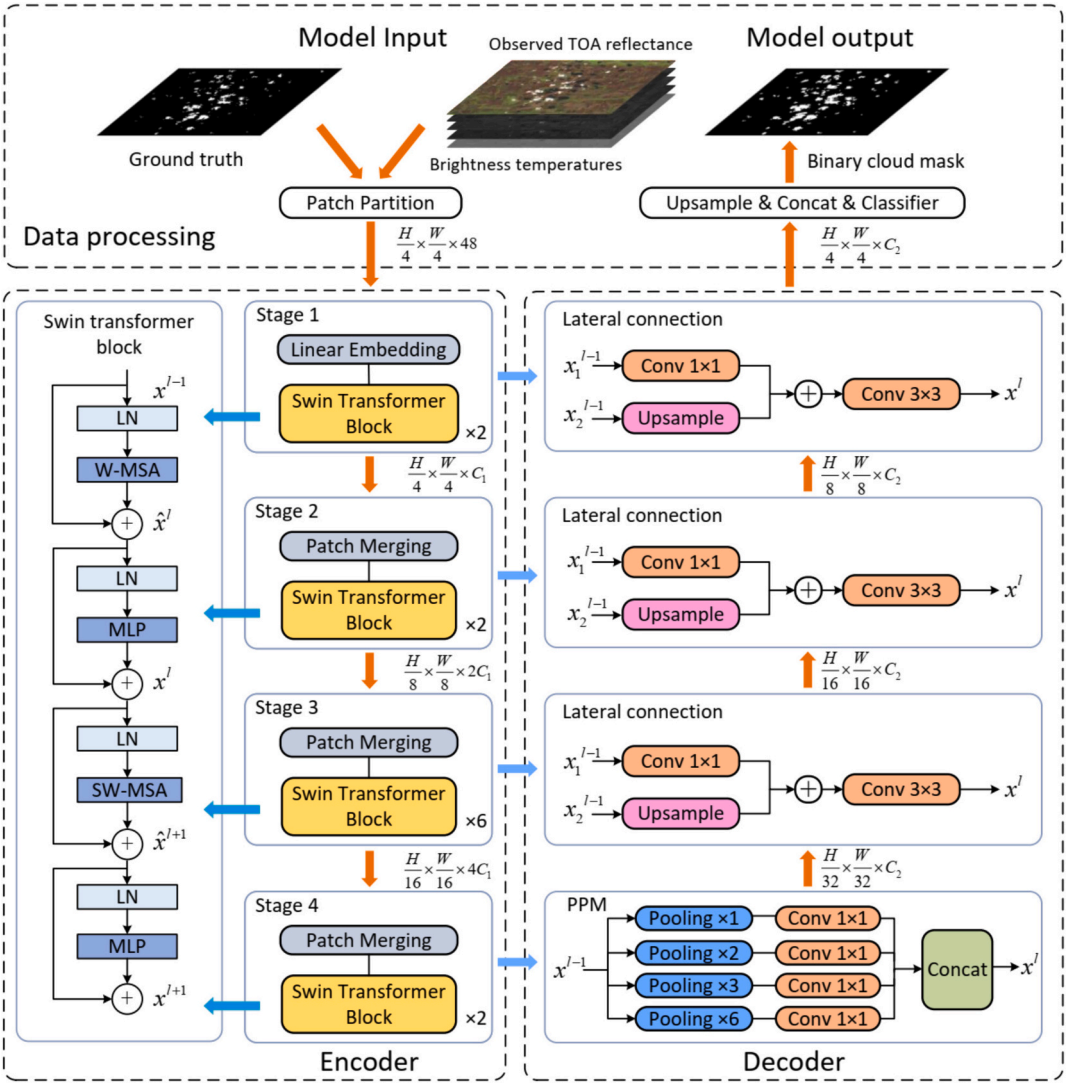

Fig. 2. Flowchart of the hybrid Swin Transformer and UPerNet (STUPmask) cloud detection model designed for satellite imagery.

or analysis (Wei et al., 2020). The remaining 48 scenes are used for validation. The patch size is set to 512 pixels with a 24-pixel overlap, resulting in 10,230 training patches and 9986 validation patches. For Sentinel-2, considering the available training sample size, the CloudSen12 training dataset, consisting of 8942 patches, each resized from the original 509 × 509 to 512 × 512 pixels (using bilinear interpolation) for model input consistency (Aybar et al., 2022). The CloudSen12 validation dataset (975 patches) and the S2 CMC dataset (2052 patches of 512 × 512 pixels and a 2-pixel overlap) are utilized for independent validation (Wright et al., 2024). The training and validation datasets are entirely independent, and all reported metrics are calculated solely based on the validation datasets. The spatial distribution of all training and validation cloud mask datasets employed in this study is shown in Fig. 1.

3. Methodology

3.1. The STUPmask framework

Regarding the limitations of traditional DL models like CNN in drawing global dependencies for satellite imagery, this study introduces the Transformer as a solution. The Transformer utilizes the self-attention mechanism to effectively capture long-range dependencies in the spatial information domain, specifically, the model's ability to capture relationships between distant pixels within the image, which is crucial for detecting clouds that span large areas or exhibit similar spatial patterns (Vaswani et al., 2017). The self-attention mechanism computes attention scores by taking the inner product of the input matrix (the image itself), followed by normalizing the attention weights using the *softmax* function. Subsequently, a weighted summation facilitates the model in effectively capturing correlations among various elements within the input sequence (Eq. 1):

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where $Q = XW^Q$, $K = XW^K$, $V = XW^V$, and d_k is the input dimension; X represents the input matrix; and W^Q, W^K and W^V denote the weight matrix. The Transformer, originally developed for Natural Language Processing (NLP), has been widely applied across various fields and has recently shown strong potential for cloud detection (Fan et al., 2024; Gong et al., 2023; Singh et al., 2023; Tan et al., 2023; B. Zhang et al., 2023; Z. Zhang et al., 2022).

The scale of the NLP problem is relatively small compared to satellite remote sensing images since processing a large volume of satellite data, especially in extracting contextual information, involves significant computational demands, as the complexity scales with the square of the number of pixels in the image. The computational complexity of pro-

cessing satellite images is proportional to the square of the number of image pixels, leading to significant computational requirements. Vision Transformer (ViT) partitions the image into several smaller blocks based on a predefined size, subsequently transforming each patch into a fixed-length vector for correlation computation (Dosovitskiy et al., 2020). However, this approach still incurs significant overhead, which is particularly noticeable when dealing with larger image sizes. The Swin Transformer provides a solution that integrates the advantages of locality, translation invariance, and hierarchy, making it suitable for image classification tasks (Liu et al., 2021). First, the hierarchical structure is introduced to process pictures, enabling the model to flexibly handle images of different scales and perform various tasks. Then, the concept of locality is introduced by applying self-attention within non-overlapping window areas, which significantly reduces computational overhead. This operation is known as Window-based Multi-head Self-Attention (W-MSA). To further facilitate information exchange between windows, the Transformer incorporates the Shifted W-MSA (SW-MSA) operation (Eq. 2):

$$\begin{aligned} \widehat{x}^{l} &= \text{W-MSA}\left(\text{LN}\left(x^{l-1}\right)\right) + x^{l-1} \\ x^{l} &= \text{MLP}\left(\text{LN}\left(\widehat{x}^{l}\right)\right) + \widehat{x}^{l} \\ \widehat{x}^{l+1} &= \text{SW-MSA}\left(\text{LN}\left(x^{l}\right)\right) + x^{l} \\ x^{l+1} &= \text{MLP}\left(\text{LN}\left(\widehat{x}^{l+1}\right)\right) + \widehat{x}^{l+1} \end{aligned}, \tag{2}$$

where LN represents the layer normalization; $\widehat{x}^{l}$ and $\widehat{x}^{l+1}$ represent the output of W-MSA and SW-MSA, respectively; x^{l} and x^{l+1} denote the output after it has passed through the multilayer perceptron (MLP) layer. This advanced design makes the Swin Transformer particularly well-suited for processing large satellite remote sensing images.

Despite the considerable attention garnered by its strong global modeling capabilities, Transformer typically treats images as a sequence of patches, potentially overlooking crucial structural information inherent in the images. In addition, merely encoding marked image blocks using a Transformer and directly upsampling the hidden features to full resolution for output often fails to yield optimal results (Chen et al., 2021; Zhang et al., 2021). The CNN architecture offers an effective approach to capturing low-level visual cues, and in this study, we chose the CNN-derived newly powerful UPerNet (Xiao et al., 2018), which features a powerful decoder that is a multi-task model capable of concurrently discerning the texture and surface attributes of objects and their diverse components. The UPerNet model uses a Pyramid Pooling Module (PPM) and lateral connections to integrate both low- and high-level information. The model structure can be adapted accordingly to achieve varying degrees of enhancement. For the current image segmentation task, feature fusion is performed at each layer of the UPerNet model, and the cross-entropy formula is employed to calculate the fused feature loss.

Here, we developed a hybrid model comprising the Swin Transformer and UPerNet models, defined as STUPmask, for satellite cloud detection (Fig. 2). The Swin Transformer serves as an encoder to enhance overall contextual understanding in satellite remote sensing images, while the decoder employs the UPerNet structure to integrate low- and high-level semantic features for cloud detection. This newly integrated design preserves the global context information extracted by the Swin Transformer and leverages the high-resolution feature map generated in the decoding path to produce a more accurate cloud classification outcome.

3.2. Model training and construction

A key aspect of the STUPmask model lies in the selection of input features for training and data collection. Imagery is first converted from digital counts to Top-of-Atmosphere (TOA) reflectance (from visible to SWIR bands) and brightness temperature (BT, for TIRS bands). The TOA reflectance of clouds notably exceeds that of common terrestrial elements in visible channels, such as water bodies, soil, vegetation, man-made structures, and rocks. In addition to visible channels, NIR and SWIR channels can also enhance cloud detection capabilities. Although the spectral characteristics of ice and snow resemble those of clouds across the visible to SWIR bands, the thermal infrared channel plays a crucial role in their differentiation due to significant differences in BTs. Furthermore, both Landsat 8 and Sentinel-2 satellites are equipped with an additional cirrus channel, which has often been used for detecting cirrus clouds (Gao et al., 2002; Zhu et al., 2015). Thus, the chosen fundamental spectral features for Landsat 8 imagery encompass visible channels spanning blue, green, red, and NIR alongside SWIR, Cirrus, and BT channels. Similarly, for the Sentinel-2 satellite, the selected basic spectral features range from blue to SWIR wavelengths, with the TIRS channel excluded due to its absence.

During the model training stage, all input parameters for the STUPmask model are first standardized. Subsequently, the AdamW optimizer is utilized with the parameters configured (e.g., $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\lambda = 0.01$) (Loshchilov and Hutter, 2017). The learning rate, determined using the Warmup strategy, begins with an initial small value and gradually increases over iterations until reaching the pre-set maximum before decay, effectively mitigating instability arising from initializing model parameters. Due to hardware memory constraints, block processing (i.e., dividing large satellite images into smaller patches and stitching the results back together) is necessary, as processing large remote-sensing images at the same time is impractical. In our study, we employed a single unified model, which was trained on a combined dataset integrating both the Landsat 8 Biome and the Sentinel-2 Cloud Mask Catalogue.

3.3. Evaluation indices

Our study employed a range of evaluation indices to quantitatively assess the STUPmask model performance. Initially, we calculated the cloud amount (CA), which refers to the cloud fraction and is defined as the ratio of cloud pixels to the total number of valid pixels in an image. We also calculated the cloud amount difference (CAD) to quantify the detected cloud content and estimate biases between the predicted and reference cloud masks generated automatically with human supervision (Foga et al., 2017). Furthermore, we computed the confusion matrix to assess the classification accuracy, indicated by the overall accuracy (OA), balanced overall accuracy (BOA), user's accuracy (UA), and producer's accuracy (PA) (Eqs. 3–6). Moreover, two indicators—F_1-score and intersection over union (IoU)—are also employed (Eqs. 7–8), where the former is the harmonic average of PA and UA, and the latter represents the intersection of two regions divided by the union of two regions. These metrics are computed based on true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP)—where FP and FN signify correctly identified cloud and non-cloud pixels, while FN and FP represent mistakenly classified cloud and non-cloud pixels.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$UA = \frac{TP}{TP + FP} \tag{4}$$

$$PA = \frac{TP}{TP + FN} \tag{5}$$

$$BOA = 0.5\left(PA + \frac{TN}{TN + FP}\right) \tag{6}$$

$$F_1 - score = \frac{2{*}UA{*}PA}{UA + PA} \tag{7}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{8}$$

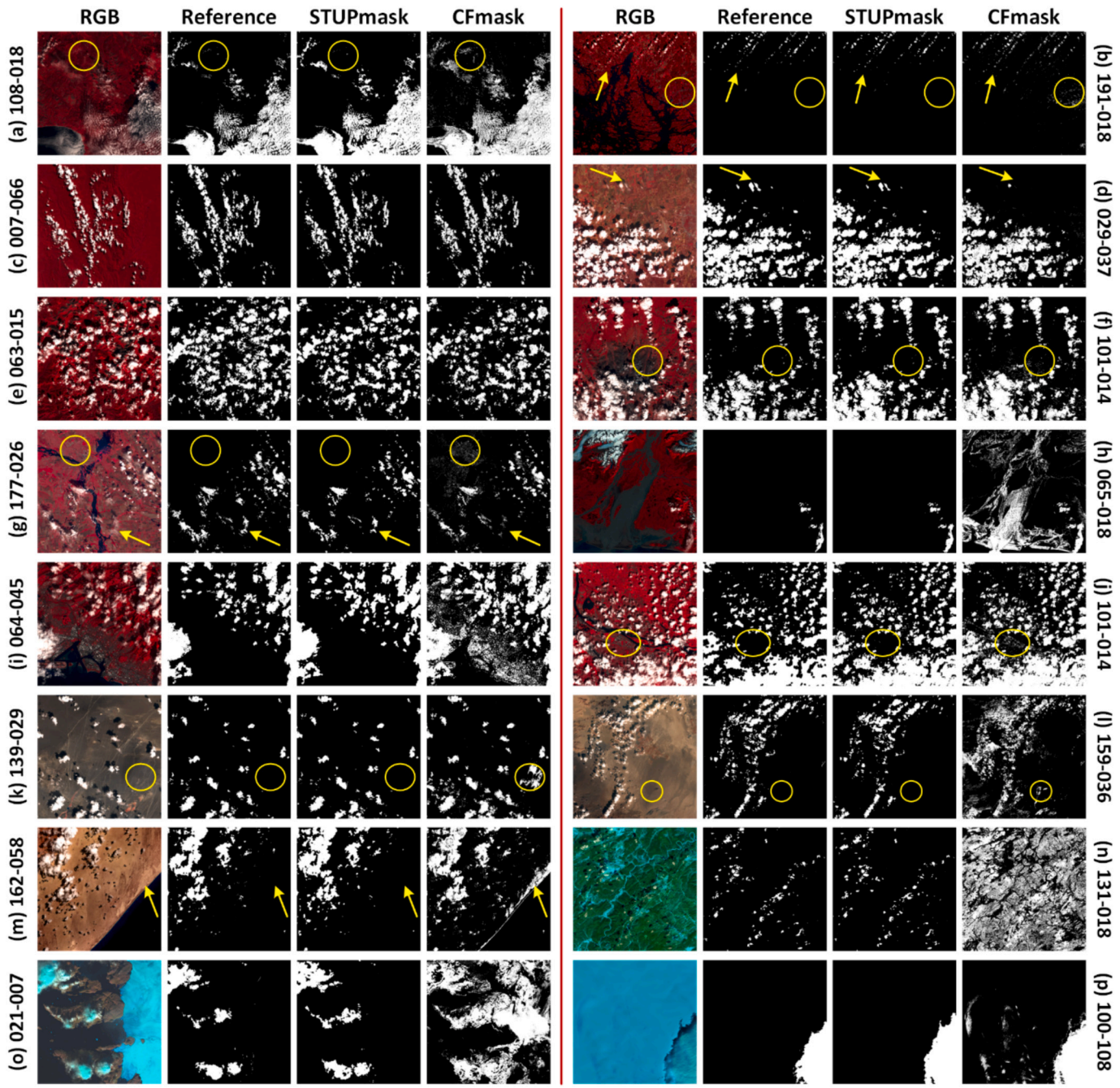

Fig. 3. Representative examples of color composite images (RGB: Bands 5–4-3 for a-m; Bands 6–5-4 for n-p to highlight snow/ice surfaces), reference cloud masks, STUPmask cloud detection results, and official CFmask cloud masks for Landsat 8 imagery (30 m). Clouds and underlying surfaces are denoted by yellow arrows and yellow ellipses, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Results and discussion

4.1. Cloud detection using the STUPmask model

4.1.1. Qualitative evaluation of Landsat 8 cloud detection

The cloud detection results for Landsat 8 Biome imagery (30 m) using the STUPmask model and the official Landsat 8 algorithm demonstrate similar spatial patterns and a high degree of consistency with the reference cloud distribution over different surfaces (Fig. 3). Overall, clouds identified by our STUPmask model and Landsat 8 official (CFmask) algorithm bear high similarity in spatial patterns and a high degree of consistency with the reference cloud distribution over dark surfaces. For example, both excel in detecting various clouds over water bodies and coastal areas (Fig. 3a, b) and are particularly adept at identifying broken clouds (pointed to by the yellow arrows), leveraging substantial differences in reflectance. Moreover, STUPmask identifies clouds well in areas with diverse vegetation features, like densely vegetated forests (Fig. 3c), agricultural regions (Fig. 3d), mountainous terrain with notable elevation variations (Fig. 3e), and even in areas with minimal cloud cover (pointed to by the yellow arrows in Fig. 3d). Furthermore, STUPmask performs well in portraying cloud distributions over areas with reduced vegetation, particularly in the mountains (Fig. 3f), vegetated landscapes mixed with small urban areas (Fig. 3g), as well as estuaries and river alluvions (Fig. 3h), exhibiting high alignment with the color composite image and minimal occurrences of omissions

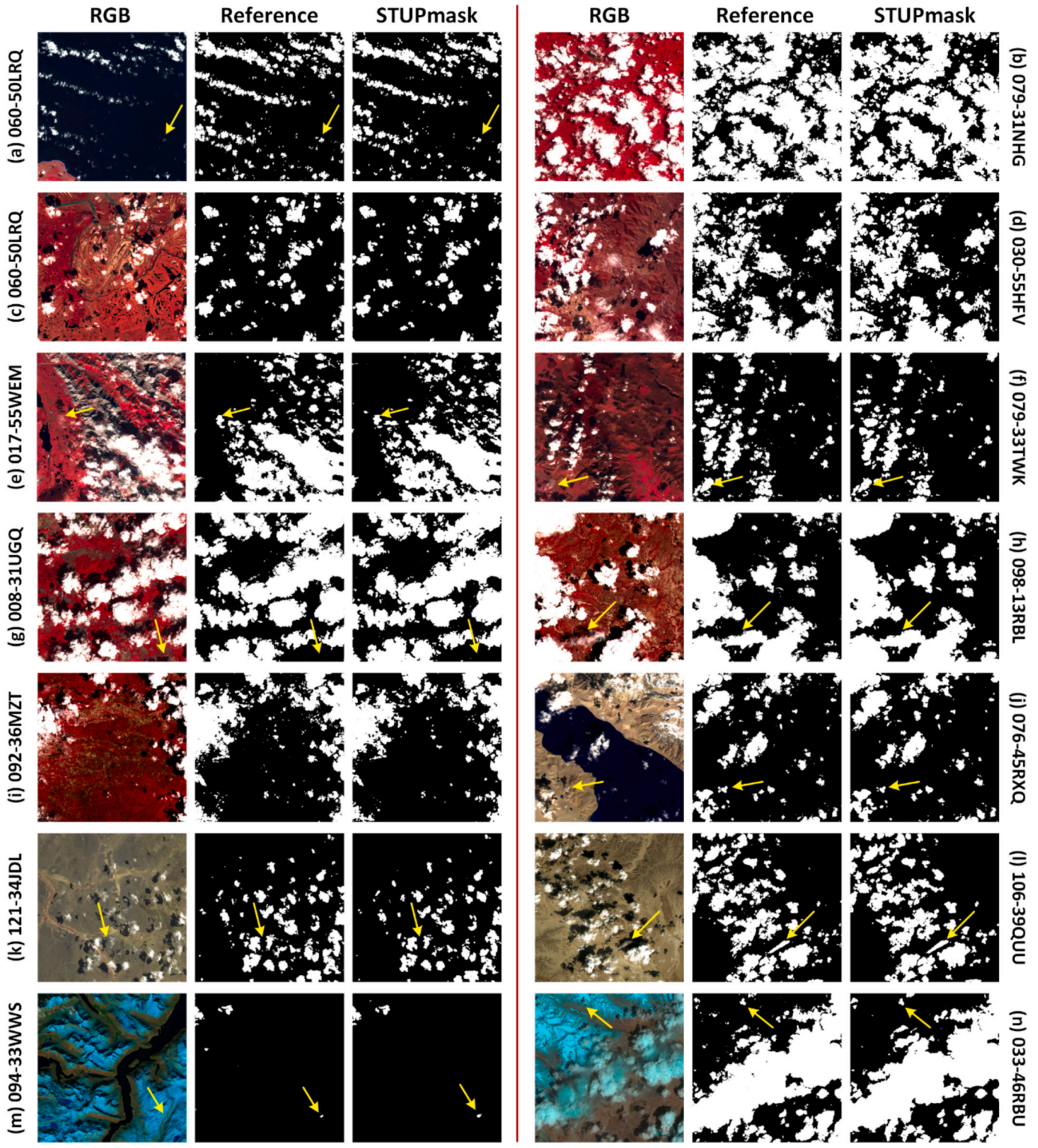

Fig. 4. Representative examples of color composite images (RGB: Bands 8–4-3 for a-l and Bands 11–8-4 for m-n to highlight snow/ice surfaces), reference cloud masks, and STUPmask cloud detection results for Sentinel-2 imagery (10 m). Clouds and underlying surfaces are denoted by yellow arrows and yellow ellipses, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and misclassifications. This performance represents significant improvements compared to the official CFmask cloud masks, as evidenced by the yellow ellipses. For instance, it is prone to misidentifying a large number of highlighted surface pixels as clouds, including man-made features like urban buildings and roads (Fig. 3i, j), bare lands like the Gobi Desert and rocks (Fig. 3k, l), and coastal tides (Fig. 3m). Additionally, it faces great challenges in identifying clouds against the backdrop of ice and snow surfaces, resulting in misjudgments (Fig. 3n–p). This is primarily attributed to the high spectral similarity between bright surfaces and clouds for traditional threshold methods, which can result in misclassifying bright surfaces as clouds and missing thin clouds.

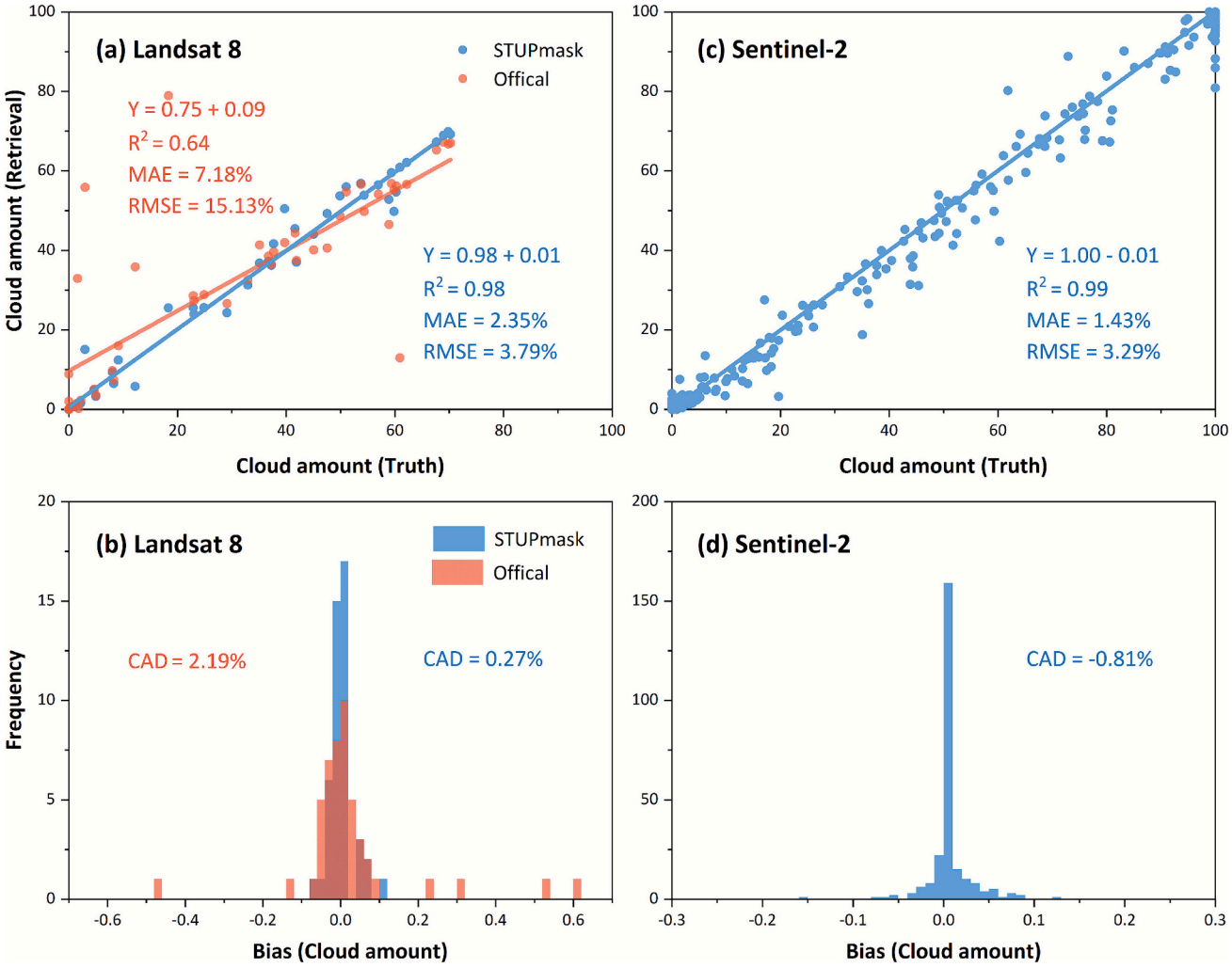

Fig. 5. Scatter plots and frequency histograms of estimated cloud amount for each image, compared with the STUPmask results with the reference cloud masks from Landsat 8 and Sentinel-2 images. The comparison with Landsat 8 official cloud mask products is also shown (indicated by orange lines and columns in a and b).

4.1.2. *Qualitative evaluation of Sentinel-2 cloud detection*

Similarly, the cloud detection results for Sentinel-2 imagery (10 m) using the STUPmask model show close spatial patterns with the cloud distribution from reference masks over different surfaces (Fig. 4). Unlike Landsat imagery, there are no official cloud mask products generated from the traditional threshold method for comparison. Our results illustrate that most clouds over the ocean (water bodies) are detected by the STUPmask model due to their large reflectance differences (Fig. 4a). Additionally, our model is effective in detecting clouds over densely vegetated and mixed-vegetated areas (Fig. 4b, c), with a high degree of consistency with the reference cloud distribution. Furthermore, the STUPmask model performs well in regions characterized by sparse vegetation, such as mountain ridge areas, where it identifies most thin and broken clouds without any further misjudgments observed (Fig. 4d, e). Our model has also been tested in challenging scenes of varying brightness. For example, the model successfully detects clouds over urban centers mixed with vegetation (indicated by yellow arrows in Fig. 4f, g), bare rocks (Fig. 4h, i), as well as the Gobi Desert and other arid areas (Fig. 4j–l). However, despite generally accurate recognition of most clouds above ice and snow surfaces (Fig. 4m, n), some cloudy pixels are still misidentified due to the absence of thermal infrared information for Sentinel-2. Overall, the STUPmask model effectively differentiates between different types of clouds (especially thin and broken clouds) for different satellites, minimizing misclassifications of clear-sky pixels as clouds over bright surfaces with minimal or no vegetation coverage.

4.2. *Quantitative evaluation of cloud detection*

4.2.1. *Cloud amount*

We first validate the cloud amounts by comparing our STUPmask-derived results with the L8 Biome and S2 CMC reference cloud mask datasets (Fig. 5). In addition, we make comparisons with available Landsat 8 official cloud masks. Our results for Landsat 8 show improved statistics (R^2 = 0.98, MAE = 2.35 %, RMSE = 3.79 %) compared to the official cloud mask product (R^2 = 0.64, MAE = 7.18 %, RMSE = 15.13 %) (Fig. 5a). In addition, the frequency histograms of STUPmask closely resemble a normal distribution (Fig. 5b), and approximately 95 % of Landsat 8 cloud detection results exhibit deviations of less than 0.1 %, indicating an average cloud amount difference (CAD) value of 0.27 %, compared to 2.19 % for the official product. Sentinel-2 shows similar accuracy, comparable to Landsat 8, featuring a high R^2 of 0.99, average MAE of 1.43 % and RMSE values of 3.29 % (Fig. 5c). Similarly, the CADs for most Sentinel-2 results are predominantly below 0.1 %, with an average of −0.81 %. The exceptional consistency between cloud amount recognition and reference images offers robust support for the rapid pre-screening of satellite data. This will significantly reduce the need for manual data selection from extensive datasets (e.g., selecting data based on a certain cloud amount threshold), ultimately saving users valuable time and resources.

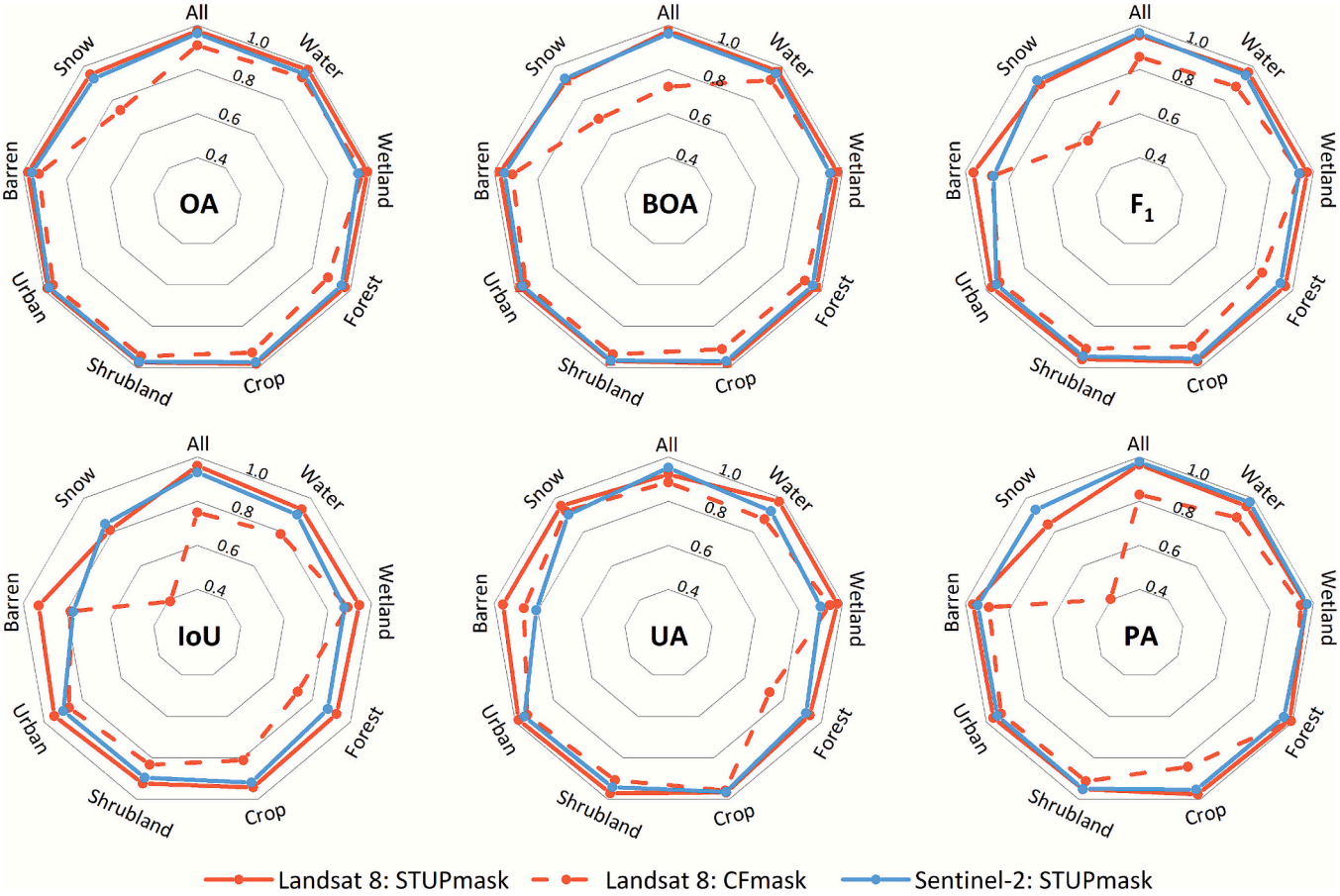

Fig. 6. Radar plots of cloud detection performance (including OA, BOA, F_1-score, IoU, UA, and PA) of our developed STUPmask model for Landsat 8 and Sentinel-2 imagery.

4.2.2. *Cloud distribution*

We then assess the STUPmask model in detecting the cloud distribution (i.e., the spatial arrangement of clouds within an image) by calculating the confusion matrix (with metrics such as PA and UA revealing typical error patterns associated with different cloud distributions) for the two satellites (Fig. 6). In general, the STUPmask model demonstrates excellent performance with Landsat 8 imagery, achieving an OA of 97.51 % and a BOA of 96.91 %, with average values for UA, PA, F_1-score, and IoU of 96.61 %, 95.32 %, 95.96 %, and 92.23 %, respectively. Notably, our new model surpasses the official algorithm, showing improvements of 7 % to 25 % in OA and BOA, and 10 % to 17 % in F_1-score and IoU, respectively. Similar excellent performance is observed for Sentinel-2 imagery using the STUPmask model, yielding an OA (BOA) of 96.27 % (96.23 %), along with a UA of 95.12 %, a PA of 97.73 %, an F_1-score of 96.41 %, and an IoU of 93.06 %. This underscores the adaptability of our model across diverse satellite sensors, ensuring precise identification of cloud distribution in remote sensing imagery. It facilitates the generation of accurate cloud masks by extracting clear-sky pixels, thereby enhancing the accuracy of remote sensing quantitative information retrieval from both the surface and atmosphere.

Furthermore, we test the model performance across different surface land cover types and found that for dark surfaces, such as shrublands (OA = 97.41 %, BOA = 96.79 %), water (OA = 98.01 %, BOA = 97.26 %), and wetlands (OA = 98.18 %, BOA = 97.78 %), the model exhibits high overall accuracy for both Landsat 8 and Sentinel-2. Additionally, the IoU scores consistently remain above 92 % for these types, underscoring that only a small portion of clouds are omitted and misclassified. STUPmask also performs well in cloud identification over urban surfaces, with an OA of 98.32 % and a BOA of 97.79 %. The model remains stable with increased reflectance of underlying surfaces, such as barren land areas, with OA (BOA) values of 97.87 % (97.48 %) and 96.11 % (95.38 %) for the two satellites. However, over snow and ice surfaces, bright surface pixels are erroneously classified as cloud pixels, resulting in a lower PA (84.50 %) compared to UA (95.57 %) for Landsat 8. In contrast, Sentinel-2 cloud recognition results exhibit fewer errors, with an average PA of 93.26 % and UA of 90.46 %. Nonetheless, the OA remains acceptable, with 95.34 % and 92.83 % for Landsat 8 and Sentinel-2, respectively. More importantly, our results consistently outperform the official cloud masks generated by the CFmask algorithm across all land use types, showing higher evaluation metrics (orange dashed lines in Fig. 6), with a particularly notable improvement over ice/snow surfaces (e.g., OA = 95.34 % *vs.* 74.11 %, BOA = 91.63 % *vs.* 69.01 %).

4.3. *Adaptable to sensors with different spatial resolutions*

In this section, we assess the transferability of our model using open-access satellite imagery from sensors with varying spatial resolutions, including meter-level data from the GaoFen-2 Panchromatic and Multispectral Sensor (PMS) (4 m) and kilometric-resolution imagery from Aqua MODIS (1 km) and the geostationary Himawari-8 Advanced Himawari Imager (AHI) (2 km). Model adaptation is accomplished by fine-tuning separately for each sensor using a limited amount of additional training data.

4.3.1. *Adaptable to very-high-resolution satellites*

First, we adapt and test our model for cloud detection using very-high-resolution (4 m) imagery from the GaoFen-2 PMS sensor using the AIR-CD dataset, which comprises 34 scenes covering diverse land

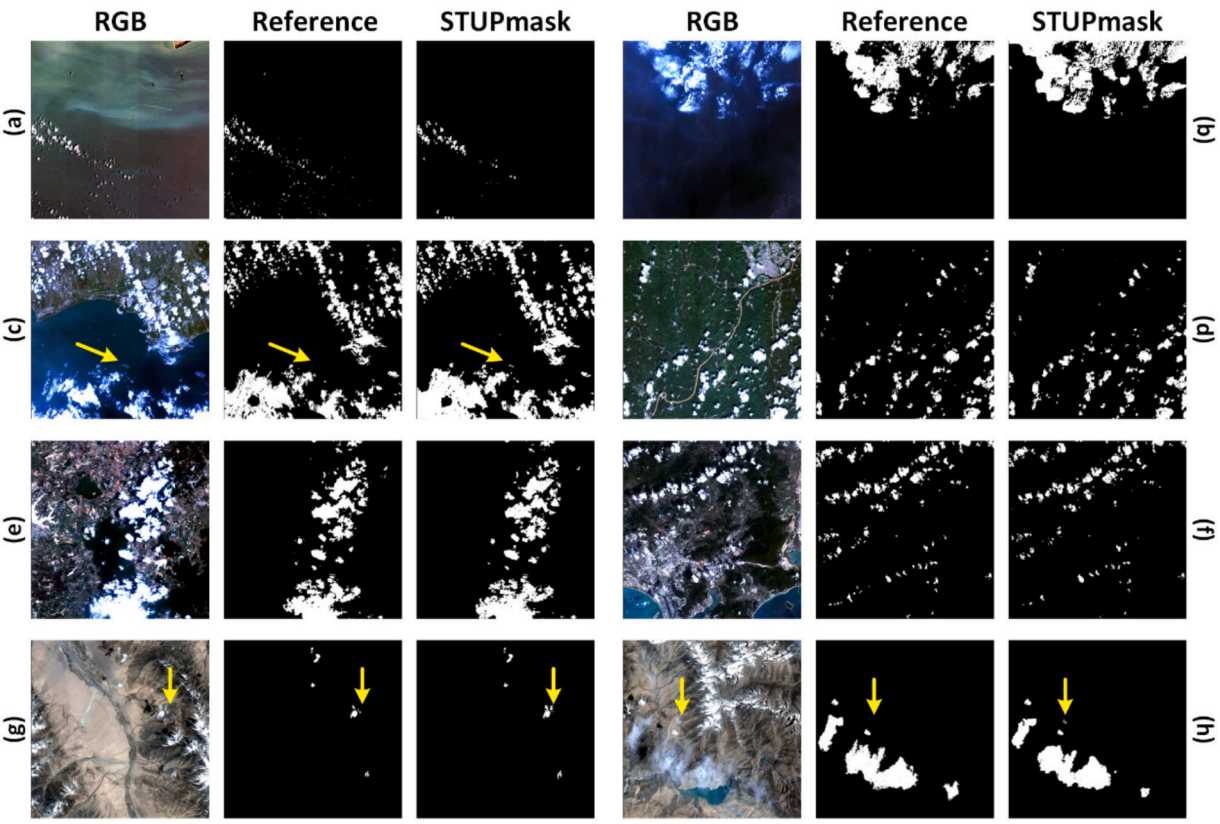

Fig. 7. Typical examples of color composite images (RGB: 3–2-1), reference cloud masks, and STUPmask cloud detection results for GaoFen-2 PMS imagery (4 m).

cover types in China (He et al., 2022). The AIR-CD dataset includes four spectral bands, spanning visible to NIR wavelengths, at a spatial resolution of 4 m and image dimensions of 7300 × 6908 pixels. For model adaptation, 20 images are randomly selected for fine-tuning, with the remaining 14 images used for validation, resulting in 4500 training samples and 3150 test samples. All other model settings remain consistent with those used for Landsat and Sentinel-2.

The cloud detection results from GaoFen-2 PMS imagery are illustrated in Fig. 7, comparing them with color composite images and reference cloud masks across different underlying surfaces. The cloud distribution classification and cloud edges are well-defined over oceans, land-water interfaces, vegetated and agricultural areas, urban buildings, barren land, and mountains. More importantly, our model performs well in distinguishing clouds without mistakenly identifying them as haze or fog (Fig. 7a, b). In vegetation-covered areas (Fig. 7c, d) and bright urban regions (Fig. 7e, f), most small, broken clouds are well-detected, while fewer clouds are observed over the rivers or lakes. Additionally, our model successfully detects thin clouds missed by the reference mask due to manual uncertainty (indicated by the yellow arrows in Fig. 7c), capturing small clouds over snow-covered high-altitude mountains (pointed to by yellow arrows in Fig. 7g, h). Even in extensive areas with thin cloud coverage, clouds are generally classified, minimizing the risk of large-scale omissions (Fig. 7h). In general, our model exhibits superior performance, with an average OA of 97.11 %, BOA of 94.46 %, F_1-score of 86.96 %, and IoU of 76.92 %, showing high PA and UA values of 91.10 % and 83.17 %, respectively, compared with the reference masks (Table S1).

4.3.2. *Adaptable to moderate-resolution satellites*

Our cloud detection model is further tested using moderate-resolution imagery from both LEO and GEO satellites, i.e., Aqua MODIS (1 km) and Himawari-8 Advanced Himawari Imager (AHI) (2 km). In this study, the MODIS cloud mask dataset includes 1272 training and 150 validation images (X. Li et al., 2022). After cropping these images into 512 × 512 patches, the final training and validation datasets consist of 19,080 and 2250 non-overlapping patches, respectively. Each patch contains ten spectral bands (1, 3, 4, 18, 20, 23, 23, 28, 29, 31, and 32), which are commonly used for cloud detection. For Himawari-8 AHI, due to the absence of manual cloud masks, the official Level 2 Cloud Mask products serve as the reference masks for this test (Takahito and Ryō, 2016). The Himawari-8 cloud mask dataset consists of 98 scenes, with 56 randomly selected for training and the remaining scenes used for validation. The images are cropped into 128 × 128 patches, yielding 20,216 training and 15,162 validation patches. Each patch contains spectral bands from visible to NIR wavelengths that are used for cloud detection. To accelerate the training process, we fine-tuned the STUPmask model and evaluated its performance using various metrics on the test dataset.

The cloud detection results from Aqua MODIS and Himawari-8 AHI imagery using the STUPmask model, compared with standard color composite images and reference cloud masks across different underlying surfaces, are shown in Fig. 8. Our model successfully detects most clouds, showing a high degree of spatial consistency with the reference cloud mask across the dark ocean (Fig. 8a, b, e, f), densely to sparsely vegetated areas (Fig. 8b, e, f), bright rock formations (Fig. 8c, g, h), and extremely bright polar ice-covered regions (Fig. 8d). More importantly, our model does not suffer from as many errors as the official cloud mask products do, particularly for small and broken clouds over bright surfaces, which are inherent limitations of multi-channel threshold-based cloud detection methods (shown by red ellipses). Nevertheless, the STUP cloud mask appears smoother and less responsive to inhomogeneity at cloud edges and holes because it is applied directly at the native resolution of the input data (e.g., 1 km for MODIS) and does not incorporate sub-pixel information from higher-resolution bands (e.g., the 250 m bands used by MODIS; Ackerman et al., 1998). Our method is designed as a conservative mask for moderate-resolution applications, which

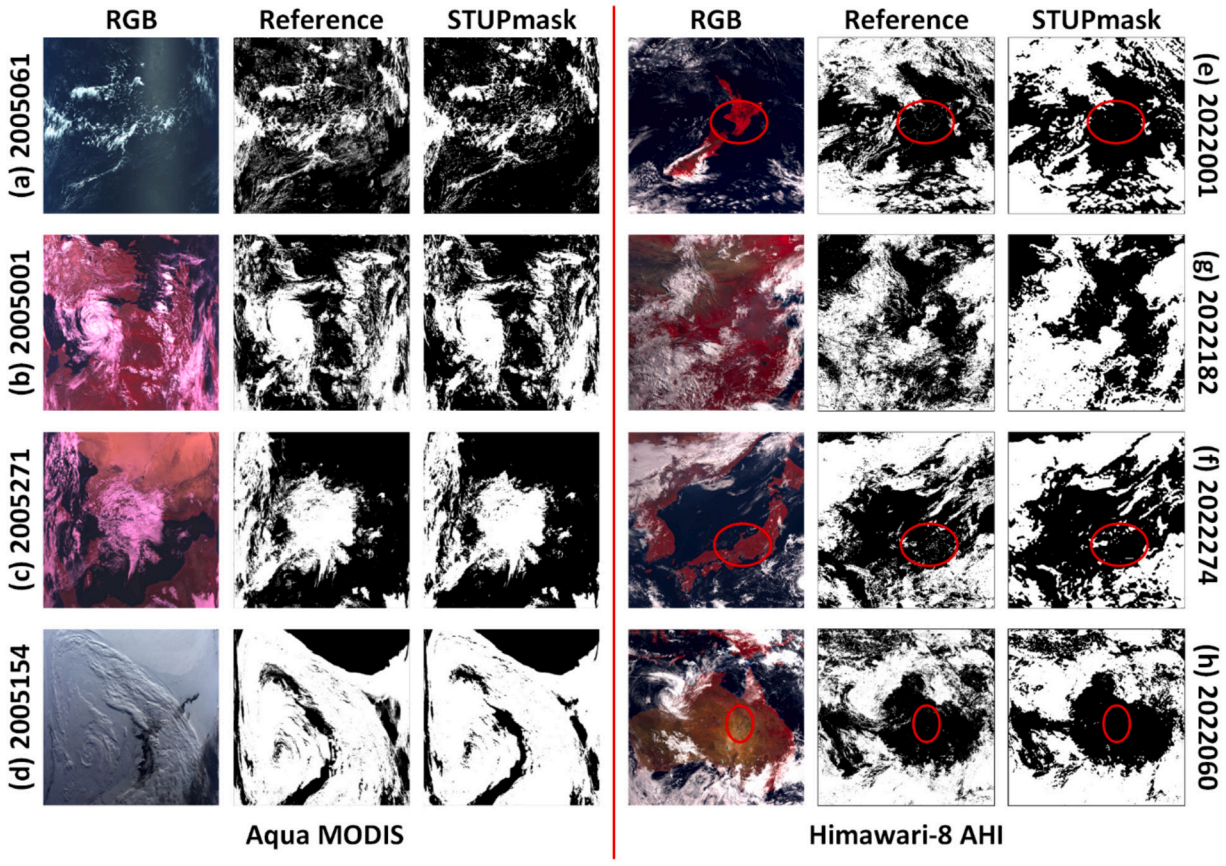

Fig. 8. Typical examples of color composite images (RGB: 2–1-4 and 4–3-2, respectively), reference cloud masks, and STUPmask cloud detection results for Aqua MODIS (1 km, left) and Himawari-8 AHI (2 km, right) imagery. The annotations on the left and right indicate the acquisition date (year followed by day number in the year) of the images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

naturally limits its ability to capture the finest-scale heterogeneity. Despite the large difference in spatial resolutions (1 km *vs.* 2 km), our model demonstrates strong visual performance in cloud detection for moderate-resolution satellite images, with no noticeable mis-classifications or missed detections. In general, our model demonstrates superior performance with OAs of 94.21 % and 96.02 %, BOAs of 94.27 % and 95.25 %, F_1-scores of 93.69 % and 91.89 %, and IoUs of 88.13 % and 84.98 % for MODIS and AHI imagery, respectively. The PA and UA values are highly comparable, with 94.84 % and 92.57 % for MODIS and 93.75 % and 90.10 % for AHI, respectively (Table S1). These results testify to the high adaptability of our STUPmask model in identifying clouds across various land cover types and satellite resolutions.

4.4. Model testing, inter-comparison, and limitations

4.4.1. Model testing and inter-comparison

First, we incorporated various publicly available datasets to comprehensively test our model and compare it with previous models, starting with L8 Spatial Procedures for Automated Removal of Cloud and Shadow (SPARCS) (80 global scenes, Fig. S1; Hughes and Kennedy, 2019) and S2 CESBIO (38 global scenes, Fig. S1; Baetens et al., 2019) cloud mask datasets, covering diverse underlying surfaces (Fig. 9). Our detected clouds show high consistency in spatial patterns with the two referenced masks. This includes clouds over oceans and inland waters, the intersection of water and land, and densely vegetated areas, with almost no obvious cloud omissions (Fig. 9a-d). In addition, our model accurately identifies fragmented or broken clouds with small amounts and particularly thin clouds covering large areas (indicated by yellow arrows in Fig. 9c, d). Superior identification results are also observed over surfaces with high reflectance, such as clouds in urban and high mountainous ice-covered areas (Fig. 9e, f), bare land (Fig. 9h), and even very challenging permanent snow/ice-covered areas (Fig. 9g), showing only a small number of clouds missed. Lastly, we quantitatively assess the accuracy of the STUPmask model using all images from the L8 SPARCS and S2 CESBIO datasets. In general, our model demonstrates superior performance with two independent validation datasets, achieving average OA and BOA values of 94.09 % and 96.42 %, 94.39 % and 94.37 % (PA = 94.81 % and 90.96 %, UA = 73.53 % and 91.01 %) for Landsat 8 and Sentinel-2 imagery compared to the conventional method. Additionally, the F_1-scores for the two satellites are 82.82 % and 90.98 %, while the IoU scores are 70.68 % and 93.47 %, respectively.

Skakun et al. (2022) conducted the Cloud Mask Intercomparison eXercise (CMIX) to benchmark 10 cloud detection algorithms, including rule-based (ATCOR, Fmask 4.0, LaSRC, Sen2Cor, Idepix), machine learning (s2cloudless, CD-FCNN), and multi-temporal approaches (FORCE, MAJA, InterSSIM) for Landsat 8 and Sentinel-2. These algorithms were implemented by the original developers under a unified evaluation framework and validated against harmonized reference masks. Following the CMIX protocol, we adopted the same evaluation procedures (Skakun et al., 2022) in our study to ensure a fair and scientifically meaningful comparison, benchmarking our model against the 10 CMIX baseline algorithms. Our model surpasses all 10 CMIX algorithms (Skakun et al., 2022), achieving the highest OA (BOA) of 96.42 % (94.37 %) on the S2 CEOBIO dataset (Fig. 10a, Table S2). When tested on the S2 PixBox dataset (17,351 pixels, Fig. S1; Paperin et al., 2021a), STUPmask excelled in detecting thin clouds, with OA and BOA reaching 91.36 % and 91.43 %, respectively, surpassing all 10 CMIX methods (Fig. 10b, Table S3). Additionally, on the L8 PixBox dataset (20,500 pixels, Fig. S1; Paperin et al., 2021b), our model demonstrated strong

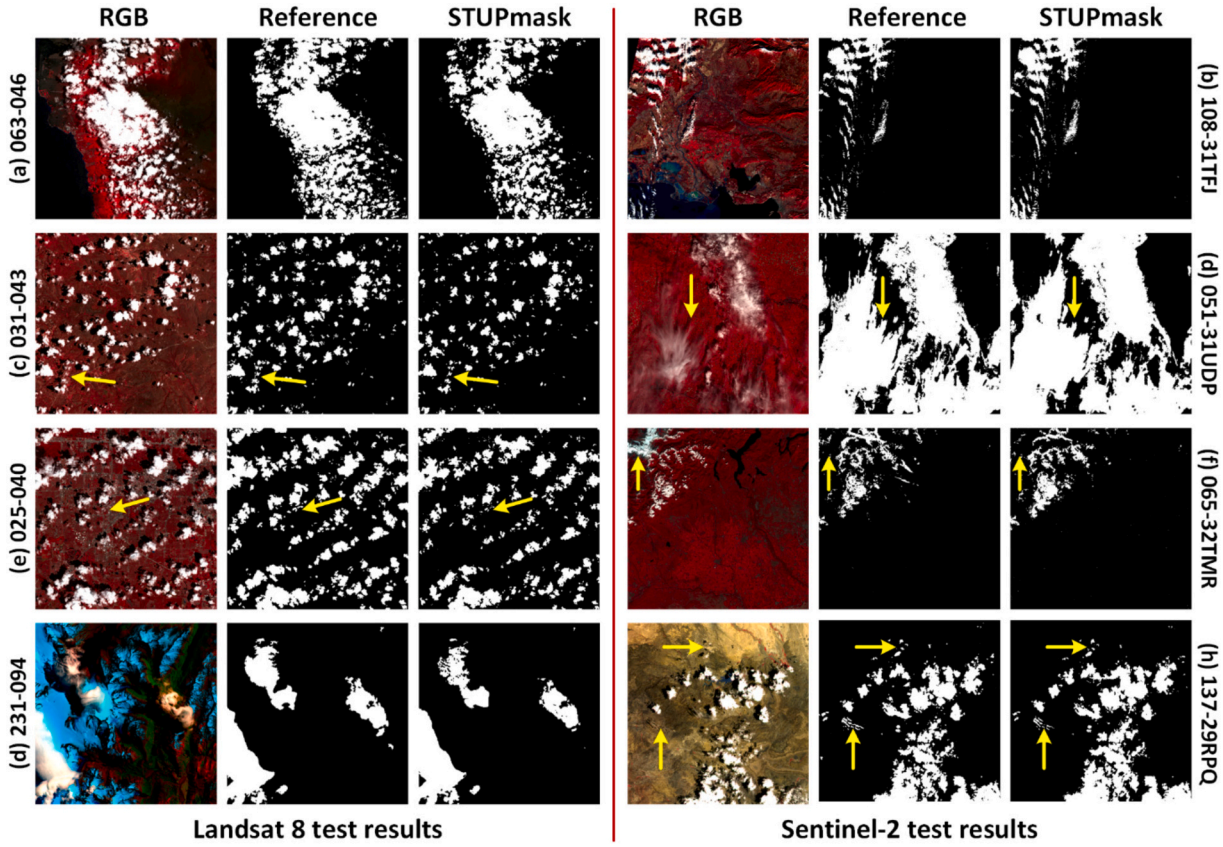

Fig. 9. Typical examples of color composite images (RGB: Bands 5–4-3 and 8–4-3, respectively), along with reference cloud masks from Landsat 8 SPARCS and Sentinel-2 CESBIO datasets and cloud detection results derived from our STUPmask model over diverse underlying surfaces.

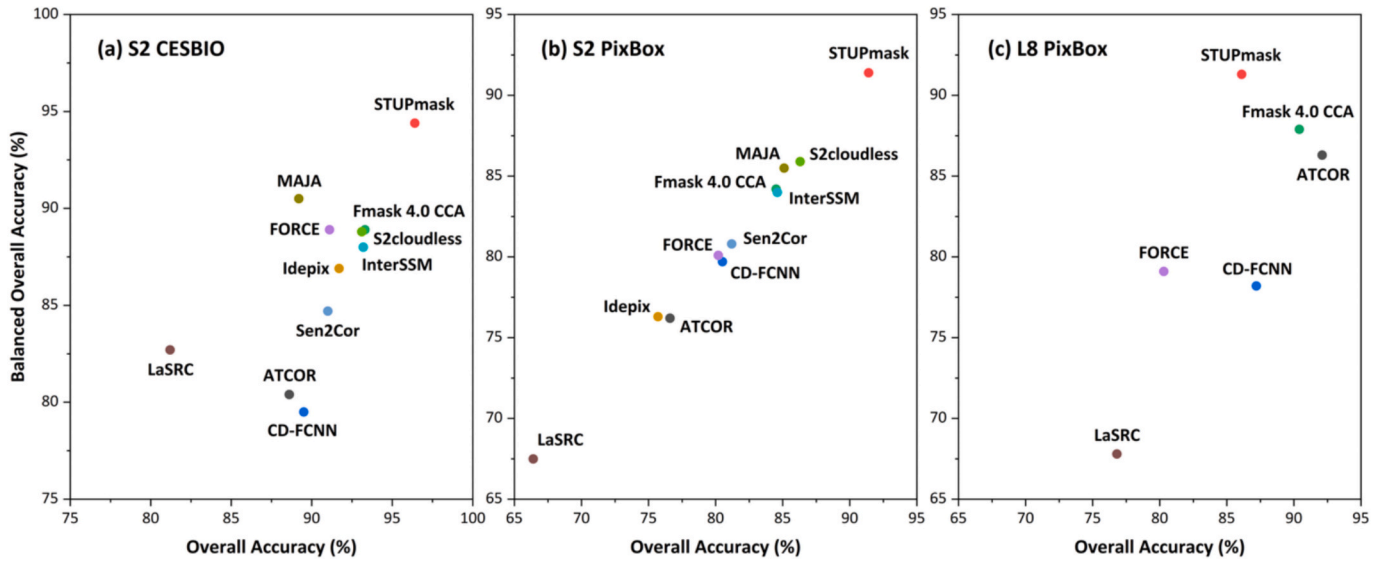

Fig. 10. Comparison of model performance between STUPmask and multiple cloud-detection algorithms from the Cloud Mask Intercomparison eXercise (CMIX) study (Skakun et al., 2022) for Sentinel-2 imagery using the (a) CESBIO and (b) PixBox datasets (all clouds), and for Landsat 8 imagery using the (c) PixBox dataset (all clouds).

performance, especially in detecting semi-transparent clouds, with the best BOA of 91.33 %, compared to 5 CMIX methods (Fig. 10c, Table S4). This improvement may be attributed to our large-scale and diverse pretraining strategy, as our training datasets, including the L8 Biome dataset, contain expert-annotated thin-cloud samples, enabling the model to learn their spectral–spatial patterns. Lastly, we evaluated our model on the S2 CloudSEN12 testing dataset (963 global scenes, Fig. S1), achieving the highest BOA of 93.64 % and outperforming all 10 benchmark models (Table S5), including DL models such as CloudS2-Mask and UNetMobV2 (Aybar et al., 2022; Wright et al., 2024). These results highlight the robust performance of our model and its applicability across diverse land and cloud conditions, surpassing widely recognized algorithms.

Table 2
Comparison of model performance with and without pretraining across different satellite sensors.

Satellite	Method	OA (%)	UA (%)	PA (%)	F_1 (%)	IoU (%)	BOA (%)
Gaofen-2	With pretraining	97.11	83.17	91.10	86.96	76.92	94.46
	Without pretraining	95.91	81.65	82.76	82.20	69.78	90.18
Himawari-8	With pretraining	96.02	90.10	93.75	91.89	84.98	95.25
	Without pretraining	95.51	89.12	91.93	90.51	82.66	94.27
MODIS	With pretraining	94.21	92.57	94.84	93.96	88.13	94.27
	Without pretraining	91.57	91.06	95.51	93.23	87.32	90.48
	MYD35	87.97	99.99	84.13	91.37	84.13	92.06

4.4.2. *Enhancing cloud detection with pretraining*

To evaluate generalization, we trained the STUPmask model from scratch on Gaofen-2, Himawari-8, and MODIS data and compared it with pre-trained models (Table 2). The pre-trained model, originally trained on Landsat and Sentinel, consistently outperforms scratch-trained models across all three sensors, achieving higher OA, F1-score, and BOA values. The pre-trained models also improve cloud detection accuracy (UA) on Gaofen-2 and Himawari-8, with a 10 % increase in completeness (PA) over the MYD35 product, and provide a more balanced precision-recall trade-off on MODIS. Models trained from scratch often misclassify bright surfaces and snow-covered areas as clouds (Fig. S2a-b), whereas pre-trained models correctly identify these regions as clear skies. This improvement results from enhanced feature discrimination through pretraining, which reduces misclassification under limited spectral information. For scenes with extensive thin or scattered clouds (Fig. S2c-e), scratch-trained models show large omission errors, while pre-trained models better capture cloud texture and morphology. Pre-trained models also preserve cloud structural continuity, whereas scratch-trained models often misclassify cloud boundaries, especially at cloud connection regions (Fig. S2f).

4.4.3. *Potential limitations*

Despite the strong performance of our model, certain limitations are worth noting. Some problematic cloud detection results obtained by the STUPmask model are presented in Fig. S3. For instance, the model struggles with low-contrast water bodies (Fig. S3a, b), where thin and semi-transparent clouds above water surfaces are difficult to identify due to minimal reflectance differences (only 0.003 in the red band). Additionally, deteriorated performance is observed over tropical deserts and barren land (Fig. S3c, d), as well as polar ice and snow regions (Fig. S3e, f), where thin clouds are often omitted, and misclassifications occur over high-reflectance surfaces (highlighted by yellow circles in the figure) (Li and Leighton, 1991). These findings underscore the inherent challenges of detecting clouds under complex and extreme conditions, emphasizing the need for continuous model improvements in future studies.

5. Conclusions

Cloud detection remains a formidable challenge given that distinguishing clouds from various background objects is hindered by their dynamic quantities and shapes, which constantly evolve over space and time. Conventional threshold-based or ML methods encounter great difficulties for thin or broken clouds over bright surfaces, particularly for satellite sensors with high spatial resolution but limited channels, such as Landsat and Sentinel. This study introduces a hybrid semantic segmentation model, named "STUPmask", which integrates the Swin Transformer and UPerNet encoder-decoder models. By combining these two approaches, the model effectively captures spatial local information of different types of clouds. We utilize radiometrically calibrated TOA reflectance and BT spanning from visible to thermal infrared bands as model inputs. The STUPmask model is trained and validated using the Landsat 8 and Sentinel-2 cloud mask validation datasets, encompassing various underlying surfaces covering the whole globe.

Our method achieves superior performance in cloud detection for both sets of satellite images compared to previous methods. The estimated cloud amounts agree well with manually annotated cloud masks, yielding R^2 values of 0.98 and 0.99, and RMSEs of 3.79 % and 3.29 %. In addition, the detected cloud distribution patterns match closely with those of the referenced masks, achieving high overall accuracies (OA) of 97.51 % and 96.27 %, and balanced OA (BOA) of 96.91 % and 96.23 %, respectively. We further test our model using various publicly available independent datasets, including Landsat 8 SPARCS, Sentinel-2 CloudSEN12 and CESBIO, as well as PixBox for both Sentinel-2 and Landsat 8, which demonstrates generally superior performance, with high OA (BOA) values ranging from 86.12 % (91.33 %) to 96.42 % (94.37 %). Our STUPmask model surpasses widely recognized models, including traditional threshold-based algorithms like the Landsat 8 official CFmask and CMIX multiple cloud mask algorithms. Relative to existing methods, the new model is particularly robust in detecting diverse cloud conditions while minimizing misclassifications over bright surfaces like ice/snow. More importantly, our model has been initially tested and applied to different LEO and GEO satellites with varying spatial resolutions (4 m–2 km), both high and low, achieving considerable accuracy (e.g., OA = 94–98 %, BOA = 94–96 %). In particular, pre-trained models demonstrate superior quantitative performance across diverse sensors and visually reduce misclassifications of bright surfaces and omissions of thin clouds, preserving cloud structures more effectively than models trained from scratch. This holds significant implications for quantitative applications in future terrestrial and atmospheric applications across a wider range of sensors with different spectral channels, which require retuning algorithms—a process that is time-consuming with conventional cloud identification methods.

CRediT authorship contribution statement

Shulin Pang: Writing – original draft, Validation, Software, Formal analysis, Data curation. **Zhanqing Li:** Writing – review & editing, Supervision. **Lin Sun:** Writing – review & editing. **Biao Cao:** Writing – review & editing, Supervision. **Zhihui Wang:** Validation, Data curation. **Xinyuan Xi:** Validation, Data curation. **Xiaohang Shi:** Validation, Data curation. **Jing Xu:** Validation, Data curation. **Jing Wei:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (42030606 and 42271412).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2025.115206.

Data availability

The links for Landsat 8 Biome, SPARCS, and PixBox datasets are https://landsat.usgs.gov/landsat-8-cloud-cover-assessment-validation-data, https://landsat.usgs.gov/cloud-validation/sparcs/l8cloudmasks.zip, and https://zenodo.org/records/5040271, and the links for the Sentinel-2 CloudSen12, CMC, CESBIO, and PixBox datasets are https://zenodo.org/records/7320147, https://zenodo.org/record/4172871, https://zenodo.org/records/1460961, and https://zenodo.org/records/5036991.

References

Ackerman, S.A., Strabala, K.I., Menzel, W.P., Frey, R.A., Moeller, C.C., Gumley, L.E., 1998. Discriminating clear sky from clouds with MODIS. J. Geophys. Res. Atmos. 103, 32141–32157. https://doi.org/10.1029/1998JD200032.

Arvidson, T., Gasch, J., Goward, S.N., 2001. Landsat 7's long-term acquisition plan - an innovative approach to building a global imagery archive. Remote Sens. Environ. 78, 13–26. https://doi.org/10.1016/s0034-4257(01)00263-2.

Asner, G.P., 2001. Cloud cover in Landsat observations of the Brazilian Amazon. Int. J. Remote Sens. 22, 3855–3862. https://doi.org/10.1080/01431160010006926.

Aybar, C., Ysuhuaylas, L., Loja, J., Gonzales, K., Herrera, F., Bautista, L., Yali, R., Flores, A., Diaz, L., Cuenca, N., Espinoza, W., Prudencio, F., Llactayo, V., Montero, D., Sudmanns, M., Tiede, D., Mateo-García, G., Gómez-Chova, L., 2022. CloudSEN12, a global dataset for semantic understanding of cloud and cloud shadow in Sentinel-2. Sci. Data 9, 782. https://doi.org/10.1038/s41597-022-01878-2.

Baetens, L., Desjardins, C., Hagolle, O., 2019. Validation of Copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure. Remote Sens. 11 (4), 433. https://doi.org/10.3390/rs11040433.

Chai, D., Newsam, S., Zhang, H.K., Qiu, Y., Huang, J., 2019. Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. Remote Sens. Environ. 225, 307–316. https://doi.org/10.1016/j.rse.2019.03.007.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Zhou, Y., 2021. TransUnet: Transformers make strong encoders for medical image segmentation. arXiv. https://doi.org/10.48550/arXiv.2102.04306.

Cheng, G., Zhou, P., Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans. Geosci. Remote Sens. 54, 7405–7415. https://doi.org/10.1109/TGRS.2016.2601622.

Deng, Z., Sun, H., Zhou, S., Zhao, J., Lei, L., Zou, H., 2018. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 145, 3–22. https://doi.org/10.1016/j.isprsjprs.2018.04.003.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. https://doi.org/10.48550/arXiv.2010.11929.

Fan, S., Song, T., Jin, G., Jin, J., Li, Q., Xia, X., 2024. A lightweight cloud and cloud shadow detection transformer with prior-knowledge guidance. IEEE Geosci. Remote Sens. Lett. 21, 1–5. https://doi.org/10.1109/LGRS.2024.3435531.

Fernández-Manso, A., Fernández-Manso, O., Quintano, C., 2016. SENTINEL-2A red-edge spectral indices suitability for discriminating burn severity. Int. J. Appl. Earth Obs. Geoinf. 50, 170–175. https://doi.org/10.1016/j.jag.2016.03.005.

Foga, S., Scaramuzza, P.L., Guo, S., Zhu, Z., Dilley, R.D., Beckmann, T., Schmidt, G.L., Dwyer, J.L., Joseph Hughes, M., Laue, B., 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. Remote Sens. Environ. 194, 379–390. https://doi.org/10.1016/j.rse.2017.03.026.

Francis, A., Mrziglod, J., Sidiropoulos, P., Muller, J.-P., 2020. Sentinel-2 cloud mask catalogue. Zenodo. https://doi.org/10.5281/zenodo.4172871.

Frantz, D., Röder, A., Udelhoven, T., Schmidt, M., 2015. Enhancing the detectability of clouds and their shadows in multitemporal dryland Landsat imagery: extending Fmask. IEEE Geosci. Remote Sens. Lett. 12, 1242–1246. https://doi.org/10.1109/LGRS.2015.2390673.

Frantz, D., Hass, E., Uhl, A., Stoffels, J., Hill, J., 2018. Improvement of the Fmask algorithm for Sentinel-2 images: separating clouds from bright surfaces based on parallax effects. Remote Sens. Environ. 215, 471–481. https://doi.org/10.1016/j.rse.2018.04.046.

Frey, R.A., Ackerman, S.A., Liu, Y., Strabala, K.I., Zhang, H., Key, J.R., Wang, X., 2008. Cloud detection with MODIS. Part I: improvements in the MODIS cloud mask for collection 5. J. Atmos. Ocean. Technol. 25, 1057–1072. https://doi.org/10.1175/2008JTECHA1052.1.

Gao, B., Yang, P., Han, W., Li, R., Wiscombe, W., 2002. An algorithm using visible and 1.38-μm channels to retrieve cirrus cloud reflectances from aircraft and satellite data. IEEE Trans. Geosci. Remote Sens. 40, 1659–1668. https://doi.org/10.1109/TGRS.2002.802454.

Ghasemian, N., Akhoondzadeh, M., 2018. Introducing two random Forest based methods for cloud detection in remote sensing images. Adv. Space Res. 62, 288–303. https://doi.org/10.1016/j.asr.2018.04.030.

Gómez-Chova, L., Amorós, J., Mateo-Garcia, G., Muñoz, J., Camps-Valls, G., 2017. Cloud masking and removal in remote sensing image time series. J. Appl. Remote. Sens. 11. https://doi.org/10.1117/1.JRS.11.015005, 015005.

Hagolle, O., Huc, M., Pascual, D.V., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. Remote Sens. Environ. 114, 1747–1755. https://doi.org/10.1016/j.rse.2010.03.002.

He, Q., Sun, X., Yan, Z., Fu, K., 2022. DABNet: deformable contextual and boundary-weighted network for cloud detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–16. https://doi.org/10.1109/TGRS.2020.3045474.

Hollstein, A., Segl, K., Guanter, L., Brell, M., Enesco, M., 2016. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in Sentinel-2 MSI images. Remote Sens. 8 (8), 666. https://doi.org/10.3390/rs8080666.

Hughes, M.J., Hayes, D.J., 2014. Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing. Remote Sens. 6 (6), 4907–4926. https://doi.org/10.3390/rs6064907.

Hughes, M.J., Kennedy, R., 2019. High-quality cloud masking of Landsat 8 imagery using convolutional neural networks. Remote Sens. 11 (21), 2591. https://doi.org/10.3390/rs11212591.

Irish, R., 2000. Landsat 7 automatic cloud cover assessment. Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI 4049, 348–355. https://doi.org/10.1117/12.410358.

Jeppesen, J.H., Jacobsen, R.H., Inceoglu, F., Toftegaard, T.S., 2019. A cloud detection algorithm for satellite imagery based on deep learning. Remote Sens. Environ. 229, 247–259. https://doi.org/10.1016/j.rse.2019.03.039.

Jin, S., Homer, C., Yang, L., Xian, G., Fry, J., Danielson, P., Townsend, P.A., 2013. Automated cloud and shadow detection and filling using two-date Landsat imagery in the USA. Int. J. Remote Sens. 34, 1540–1560. https://doi.org/10.1080/01431161.2012.720045.

King, M.D., Platnick, S., Menzel, W.P., Ackerman, S.A., Hubanks, P.A., 2013. Spatial and temporal distribution of clouds observed by MODIS onboard the Terra and Aqua satellites. IEEE Trans. Geosci. Remote Sens. 51, 3826–3852. https://doi.org/10.1109/tgrs.2012.2227333.

Kriebel, K.T., Saunders, R.W., Gesell, G., 1989. Optical properties of clouds derived from fully cloudy AVHRR pixels. Beiträge zur Physik der Atmosphäre 62, 165–171.

Gong, C., Long, T., Yin, R., Jiao, W., Wang, G., 2023. A hybrid algorithm with Swin transformer and convolution for cloud detection. Remote Sens. 15 (21), 5264. https://doi.org/10.3390/rs15215264.

Li, B., Liang, S., Liu, X., Ma, H., Chen, Y., Liang, T., He, T., 2021. Estimation of all-sky 1 km land surface temperature over the conterminous United States. Remote Sens. Environ. 266, 112707. https://doi.org/10.1016/j.rse.2021.112707.

Li, S., Dragicevic, S., Castro, F.A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., Cheng, T., 2016. Geospatial big data handling theory and methods: a review and research challenges. ISPRS J. Photogramm. Remote Sens. 115, 119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012.

Li, X., Yang, X., Li, X., Lu, S., Ye, Y., Ban, Y., 2022. GCDB-UNet: a novel robust cloud detection approach for remote sensing images. Knowl.-Based Syst. 238, 107890. https://doi.org/10.1016/j.knosys.2021.107890.

Li, Z., Leighton, H.G., 1991. Scene identification and its effect on cloud radiative forcing in the Arctic. J. Geophys. Res. Atmos. 96, 9175–9188. https://doi.org/10.1029/91JD00529.

Li, Z., Shen, H., Cheng, Q., Liu, Y., You, S., He, Z., 2019. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. ISPRS J. Photogramm. Remote Sens. 150, 197–212. https://doi.org/10.1016/j.isprsjprs.2019.02.017.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1711.05101.

Luo, W., Li, Y., Urtasun, R., Zemel, R., 2016. Understanding the effective receptive field in deep convolutional neural networks. Adv. Neural Inf. Proces. Syst. 29, 4898–4906. https://doi.org/10.48550/arXiv.1701.04128.

Ozkan, S., Efendioglu, M., Demirpolat, C., 2018. Cloud detection from RGB color remote sensing images with deep pyramid networks. In: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 6939–6942. https://doi.org/10.1109/IGARSS.2018.8519570.

Paperin, M., Wevers, J., Stelzer, K., Brockmann, C., 2021a. PixBox Sentinel-2 pixel collection for CMIX (version 1.0). Zenodo. https://doi.org/10.5281/zenodo.5036991.

Paperin, M., Stelzer, K., Lebreton, C., Brockmann, C., Wevers, J., 2021b. PixBox Landsat 8 pixel collection for CMIX (version 1.0). Zenodo. https://doi.org/10.5281/zenodo.5040271.

Pérez-Suay, A., Amorós-López, J., Gómez-Chova, L., Muñoz-Marí, J., Just, D., Camps-Valls, G., 2018. Pattern recognition scheme for large-scale cloud detection over landmarks. IEEE J. Select. Topics Appl. Earth Observ. Remote Sens. 11, 3977–3987. https://doi.org/10.1109/JSTARS.2018.2863383.

Qiu, S., He, B., Zhu, Z., Liao, Z., Quan, X., 2017. Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images. Remote Sens. Environ. 199, 107–119. https://doi.org/10.1016/j.rse.2017.07.002.

Qiu, S., Zhu, Z., He, B., 2019. Fmask 4.0: improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. Remote Sens. Environ. 231, 111205. https://doi.org/10.1016/j.rse.2019.05.024.

Saunders, R.W., Kriebel, K.T., 1988. An improved method for detecting clear sky and cloudy radiances from AVHRR data. Int. J. Remote Sens. 9, 123–150. https://doi.org/10.1080/01431168808954841.

Scaramuzza, P.L., Bouchard, M.A., Dwyer, J.L., 2012. Development of the Landsat data continuity mission cloud-cover assessment algorithms. IEEE Trans. Geosci. Remote Sens. 50, 1140–1154. https://doi.org/10.1109/TGRS.2011.2164087.

Schneider, A., Friedl, M.A., Potere, D., 2010. Mapping global urban areas using MODIS 500-m data: new methods and datasets based on 'urban ecoregions'. Remote Sens. Environ. 114, 1733–1746. https://doi.org/10.1016/j.rse.2010.03.003.

Singh, R., Mantosh, B., Pal, M., 2023. A transformer-based cloud detection approach using Sentinel 2 imageries. Int. J. Remote Sens. 44, 3194–3208. https://doi.org/10.1080/01431161.2023.2216850.

Skakun, S., Wevers, J., Brockmann, C., Doxani, G., Aleksandrov, M., Batič, M., Frantz, D., Gascon, F., Gómez-Chova, L., Hagolle, O., López-Puigdollers, D., Louis, J., Lubej, M., Mateo-García, G., Osman, J., Peressutti, D., Pflug, B., Puc, J., Richter, R., Roger, J.-C., Scaramuzza, P., Vermote, E., Vesel, N., Zupanc, A., Žust, L., 2022. Cloud mask Intercomparison eXercise (CMIX): an evaluation of cloud masking algorithms for Landsat 8 and Sentinel-2. Remote Sens. Environ. 274, 112990. https://doi.org/10.1016/j.rse.2022.112990.

Stowe, L.L., McClain, E.P., Carey, R., Pellegrino, P., Gutman, G.G., Davis, P., Long, C., Hart, S., 1991. Global distribution of cloud cover derived from NOAA/AVHRR operational satellite data. Adv. Space Res. 11, 51–54. https://doi.org/10.1016/0273-1177(91)90402-6.

Sui, Y., He, B., Fu, T., 2019. Energy-based cloud detection in multispectral images based on the SVM technique. Int. J. Remote Sens. 40, 5530–5543. https://doi.org/10.1080/01431161.2019.1580788.

Sun, L., Wei, J., Wang, J., Mi, X.T., Guo, Y.M., Lv, Y., Yang, Y.K., Gan, P., Zhou, X.Y., Jia, C., Tian, X.P., 2016. A universal dynamic threshold cloud detection algorithm (UDTCDA) supported by a prior surface reflectance database. J. Geophys. Res.-Atmos. 121, 7172–7196. https://doi.org/10.1002/2015jd024722.

Takahito, I., Ryō, Y., 2016. Algorithm theoretical basis for Himawari-8 cloud mask product. Meteorological Satellite Center Technical Note 61, 1–17.

Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., Brisco, B., 2020. Google earth engine for geo-big data applications: a meta-analysis and systematic review. ISPRS J. Photogramm. Remote Sens. 164, 152–170. https://doi.org/10.1016/j.isprsjprs.2020.04.001.

Tan, Y., Zhang, W., Yang, X., Liu, Q., Mi, X., Li, J., Yang, J., Gu, X., 2023. Cloud and cloud shadow detection of GF-1 images based on the Swin-UNet method. Atmosphere 14, 1669. https://doi.org/10.3390/atmos14111669.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst. 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762.

Wang, Y., Sun, Y., Cao, X., Wang, Y., Zhang, W., Cheng, X., 2023. A review of regional and global scale land use/land cover (LULC) mapping products generated from satellite remote sensing. ISPRS J. Photogramm. Remote Sens. 206, 311–334. https://doi.org/10.1016/j.isprsjprs.2023.11.014.

Wei, J., Huang, W., Li, Z., Sun, L., Zhu, X., Yuan, Q., Liu, L., Cribb, M., 2020. Cloud detection for Landsat imagery by combining the random forest and superpixels extracted via energy-driven sampling segmentation approaches. Remote Sens. Environ. 248, 112005. https://doi.org/10.1016/j.rse.2020.112005.

Wei, J., Li, Z., Lyapustin, A., Wang, J., Dubovik, O., Schwartz, J., Sun, L., Li, C., Liu, S., Zhu, T., 2023. First close insight into global daily gapless 1 km $PM_{2.5}$ pollution, variability, and health impact. Nat. Commun. 14, 8349. https://doi.org/10.1038/s41467-023-43862-3.

Wei, J., Wang, Z., Li, Z., Li, Z., Pang, S., Xi, X., Cribb, M., Sun, L., 2024. Global aerosol retrieval over land from Landsat imagery integrating transformer and Google earth engine. Remote Sens. Environ. 315, 114404. https://doi.org/10.1016/j.rse.2024.114404.

Wieland, M., Li, Y., Martinis, S., 2019. Multi-sensor cloud and cloud shadow segmentation with a convolutional neural network. Remote Sens. Environ. 230, 111203. https://doi.org/10.1016/j.rse.2019.05.022.

Wright, N., Duncan, J.M.A., Callow, J.N., Thompson, S.E., George, R.J., 2024. CloudS2Mask: a novel deep learning approach for improved cloud and cloud shadow masking in Sentinel-2 imagery. Remote Sens. Environ. 306, 114122. https://doi.org/10.1016/j.rse.2024.114122.

Wright, N., Duncan, J.M.A., Callow, J.N., Thompson, S.E., George, R.J., 2025. Training sensor-agnostic deep learning models for remote sensing: achieving state-of-the-art cloud and cloud shadow identification with OmniCloudMask. Remote Sens. Environ. 322, 114694.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. ECCV 2018, Lecture Notes in Computer Science, vol. 11209. Springer, Cham. https://doi.org/10.1007/978-3-030-01228-1_26.

Zhang, Y.C., Rossow, W.B., Lacis, A.A., Oinas, V., Mishchenko, M.I., 2004. Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: refinements of the radiative transfer model and the input data. J. Geophys. Res.-Atmos. 109, D19105. https://doi.org/10.1029/2003jd004457.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Proces. Syst. 34, 12077–12090. https://doi.org/10.48550/arXiv.2105.15203.

Zhang, Y., Liu, H., Hu, Q., 2021. TransFuse: Fusing transformers and CNNs for medical image segmentation. In: de Bruijne, M., et al. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021, Lecture Notes in Computer Science, vol. 12901. Springer, Cham. https://doi.org/10.1007/978-3-030-87193-2_2.

Zhang, Z., Xu, Z., Liu, C., Tian, Q., Wang, Y., 2022. Cloudformer: supplementary aggregation feature and mask-classification network for cloud detection. Appl. Sci. 12 (7), 3221. https://doi.org/10.3390/app12073221.

Zhang, B., Zhang, Y., Li, Y., Wan, Y., Yao, Y., 2023. CloudViT: a lightweight vision transformer network for remote sensing cloud detection. IEEE Geosci. Remote Sens. Lett. 20, 1–5. https://doi.org/10.1109/LGRS.2022.3233122.

Zhang, H.K., Luo, D., Roy, D.P., 2024. Improved Landsat operational land imager (OLI) cloud and shadow detection with the learning attention network algorithm (LANA). Remote Sens. 16 (8), 1321. https://doi.org/10.3390/rs16081321.

Zhen, Z., Chen, S., Yin, T., Gastellu-Etchegorry, J.-P., 2023. Globally quantitative analysis of the impact of atmosphere and spectral response function on 2-band enhanced vegetation index (EVI2) over Sentinel-2 and Landsat-8. ISPRS J. Photogramm. Remote Sens. 205, 206–226. https://doi.org/10.1016/j.isprsjprs.2023.09.024.

Zheng, Q., Yang, M., Yang, J., Zhang, Q., Zhang, X., 2018. Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process. IEEE Access 6, 15844–15869. https://doi.org/10.1109/ACCESS.2018.2810849.

Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028.

Zhu, Z., Woodcock, C.E., 2014. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: an algorithm designed specifically for monitoring land cover change. Remote Sens. Environ. 152, 217–234. https://doi.org/10.1016/j.rse.2014.06.012.

Zhu, Z., Wang, S.X., Woodcock, C.E., 2015. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4-7, 8, and Sentinel 2 images. Remote Sens. Environ. 159, 269–277. https://doi.org/10.1016/j.rse.2014.12.014.